Archetypal Distribution

Sohan Seth

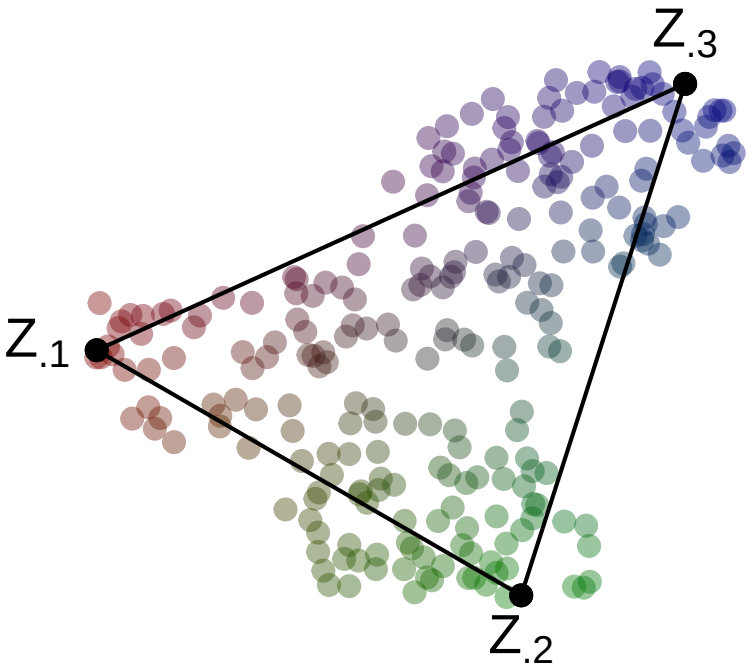
University of Edinburgh, School of Informatics, Edinburgh, EH8 9AB, UK

January 21, 2020

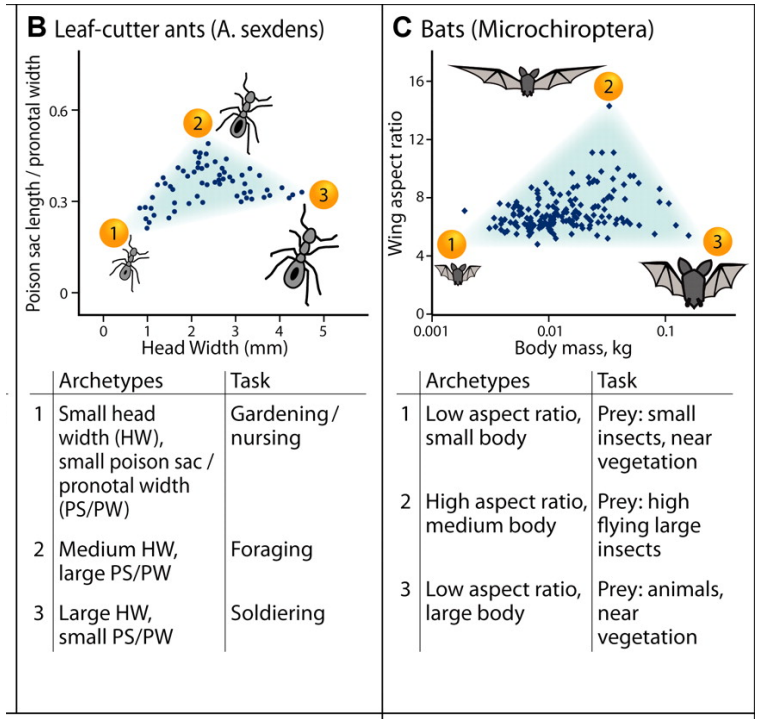
THE UNIVERSITY *of* EDINBURGH

Archetypal Analysis

- Archetypes are prototypes, i.e., representative observations, that are ideal examples of a type
- Archetypes are interpretable since they relate to actual observations
- Archetypes are extreme in nature rather than *average*, for example, as in *medoids*



Seth and Eugster [2016]



	Archetypes	Task
1	Small head width (HW), small poison sac / pronotal width (PS/PW)	Gardening / nursing
2	Medium HW, large PS/PW	Foraging
3	Large HW, small PS/PW	Soldiering

	Archetypes	Task
1	Low aspect ratio, small body	Prey: small insects, near vegetation
2	High aspect ratio, medium body	Prey: high flying large insects
3	Low aspect ratio, large body	Prey: animals, near vegetation

Shoval et al., [2012]

From Value to Distribution

A single value often does not carry all the information, e.g.,

- two scientific papers can have scores $\{3, 3, 4\}$ and $\{1, 4, 5\}$
- two movies can have ratings $\{6, 7, 9\}$ and $\{4, 9, 9\}$.

How do we find archetypes over distributions?

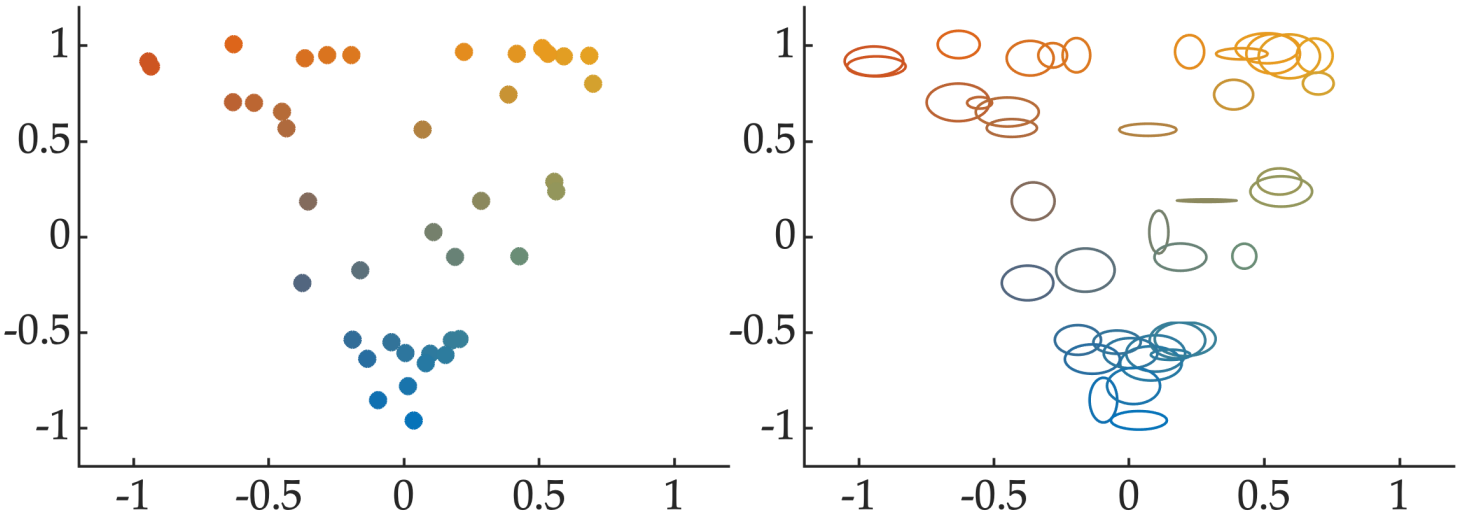


Figure: **(left)** standard archetypal analysis **(right)** archetypal analysis over distributions

Standard Archetypal Analysis [Cutler and Breiman, 1994]

Given a set of n observations $\mathbf{x}_1, \ldots, \mathbf{x}_n$,

1. Define K archetypes $\mathbf{z}_1, \ldots, \mathbf{z}_K$ as convex combinations of the observations, i.e.,

$$\mathbf{z}_k = \sum_{i=1}^{n} w_{ik} \mathbf{x}_i$$

   where $\mathbf{W} \in \Delta$ is a $n \times k$ dimensional matrix with $[\mathbf{W}]_{ik} = w_{ik}$.

2. Reconstruct the observations as convex combinations of the archetypes, i.e.,

$$\hat{\mathbf{x}}_j = \sum_{k=1}^{K} h_{kj} \mathbf{z}_k$$

   where $\mathbf{H} \in \Delta$ is a $k \times n$ dimensional matrix with $[\mathbf{H}]_{kj} = h_{kj}$.

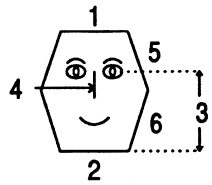
3. Optimize the parameters $\mathbf{W}$ and $\mathbf{H}$ by minimizing the error between the observations and their respective reconstructions, i.e.,

$$\sum_{i=1}^{n} ||\mathbf{x}_i - \hat{\mathbf{x}}_i||^2$$

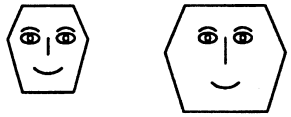
   where $||\cdot||$ denotes the l_2-norm. Thus, archetypal analysis can be summarized as

$$\min_{\mathbf{W},\mathbf{H} \in \Delta} ||\mathbf{X} - \mathbf{XWH}||^2.$$

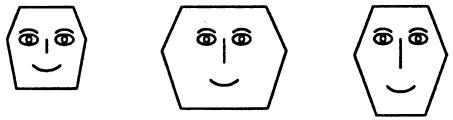
Standard Archetypal Analysis [Cutler and Breiman, 1994]



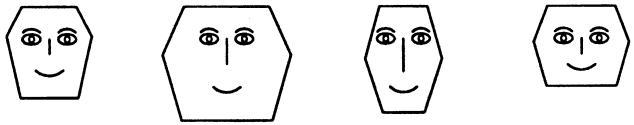
Two Archetypes



Three Archetypes



Four Archetypes



Five Archetypes

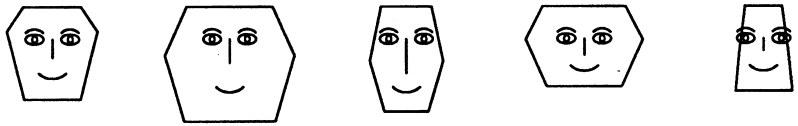


Figure 1. Archetypes for Head Dimension Data.

Kernel Archetypal Analysis [Mørup and Hansen, 2012]

Use 'kernel trick' for distributions with Bhattacharyya coefficient.

- The cost can be rewritten as

$$||\mathbf{X} - \mathbf{XWH}||^2 = \mathrm{tr}((\mathbf{I} - \mathbf{WH})^\intercal \mathbf{X}^\intercal \mathbf{X}(\mathbf{I} - \mathbf{WH}))$$

  where tr denotes the trace operation.
- The inner product $[\mathbf{X}^\intercal \mathbf{X}]_{ij} = \langle \mathbf{x}_i \,|\, \mathbf{x}_j \rangle$ can then be replaced by a positive definite kernel

$$[\mathbf{K}]_{ij} = \kappa(\mathbf{x}_i, \mathbf{x}_j) = \langle \phi(\mathbf{x}_i) \,|\, \phi(\mathbf{x}_j) \rangle,$$

  where ϕ is a (nonlinear) mapping from $\mathbb{R}^d$ to a feature space $\mathcal{F}$ with inner product κ.

Pros

- Allows performing archetypal analysis in any observation space with a p.d. kernel

Cons

- Result depends on the choice of kernel
- Not knowing the explicit mapping ϕ hinders the interpretability of the archetypes

Interval Archetypal Analysis [D'Esposito et al., 2012]

- Each element of the matrix $\mathbf{X}$ is an interval, i.e., $[\mathbf{X}]_{lj} = [\underline{x}_{lj}, \bar{x}_{lj}]$
- Use weighted sum and distance operation with appropriate operations on interval , i.e.,

$$\underline{z}_{lk} = \sum_{i=1}^{n} w_{ik}\underline{x}_{li} \text{ and } \bar{z}_{lk} = \sum_{i=1}^{n} w_{ik}\bar{x}_{li}, \text{ and}$$

$$d(x, y) = \max(|\bar{x} - \bar{y}|, |\underline{x} - \underline{y}|) = |x_m - y_m| + |x_d - y_d|$$

  where $x_m = (\underline{x} + \bar{x})/2$ and $x_d = (\bar{x} - \underline{x})/2$.
- Archetypal analysis on intervals can be summarized as

$$\min_{\mathbf{W},\mathbf{H}\in\Delta} |\mathbf{X}_m - \mathbf{X}_m\mathbf{W}\mathbf{H}| + |\mathbf{X}_d - \mathbf{X}_d\mathbf{W}\mathbf{H}|.$$

Pros

- Straighforward extension of standard archetypal analysis

Cons

- Limited to intervals not continuous distributions

Interval Archetypal Analysis [D'Esposito et al., 2012]

Table 1. Bats data set [26].

i	Species	Head	Tail	Height	Forearm
1	PIPC	33, 52	26, 33	4, 7	27, 32
2	PRH	35, 43	24, 30	8, 11	34, 41
3	MOUS	38, 50	30, 40	7, 8	32, 37
4	PIPS	43, 48	34, 39	6, 7	31, 38
5	PIPN	44, 48	34, 44	7, 8	31, 36
6	MDAUB	41, 51	30, 39	8, 11	33, 41
7	MNAT	42, 50	32, 43	8, 9	36, 42
8	MDEC	40, 45	39, 44	9, 9	36, 42
9	MGP	45, 53	35, 38	10, 12	39, 44
10	OCOM	41, 51	34, 50	9, 10	34, 50
11	MBEC	46, 53	34, 44	9, 11	39, 44
12	SBOR	48, 54	38, 47	9, 11	37, 42
13	BARB	44, 58	41, 54	6, 8	35, 41
14	OGRIS	47, 53	43, 53	7, 9	37, 41
15	SBIC	50, 63	40, 45	8, 10	40, 47
16	FCHEV	50, 69	30, 43	11, 13	51, 61
17	MSCH	52, 60	50, 60	10, 11	42, 48
18	SCOM	62, 80	46, 57	9, 12	48, 56
19	NOCT	69, 82	41, 59	10, 12	45, 55
20	GMUR	65, 80	48, 60	12, 16	55, 68
21	MGES	82, 87	46, 57	11, 12	58, 63

Statistical Archetypal Analysis [Wu and Tabak, 2017]

- Finds archetypal distribution as mixture of observed distribution, i.e.,

$$q_k(\mathbf{x}) = \sum_{i=1}^{n} w_{ik} p_i(\mathbf{x}),$$

- Minimize the 'energy distance' between the observed and reconstructed distributions

$$\hat{p}_j(\mathbf{x}) = \sum_{k=1}^{K} h_{kj} q_k(\mathbf{x})$$

.

- Energy distance between two distributions p and q is defined as

$$\mathrm{D}_{\mathrm{ED}}(p,q) = -\mathbb{E}_{X,X'\sim p}||X-X'|| - \mathbb{E}_{Y,Y'\sim q}||Y-Y'|| + 2\mathbb{E}_{X\sim p,Y\sim q}||X-Y||.$$

Pros

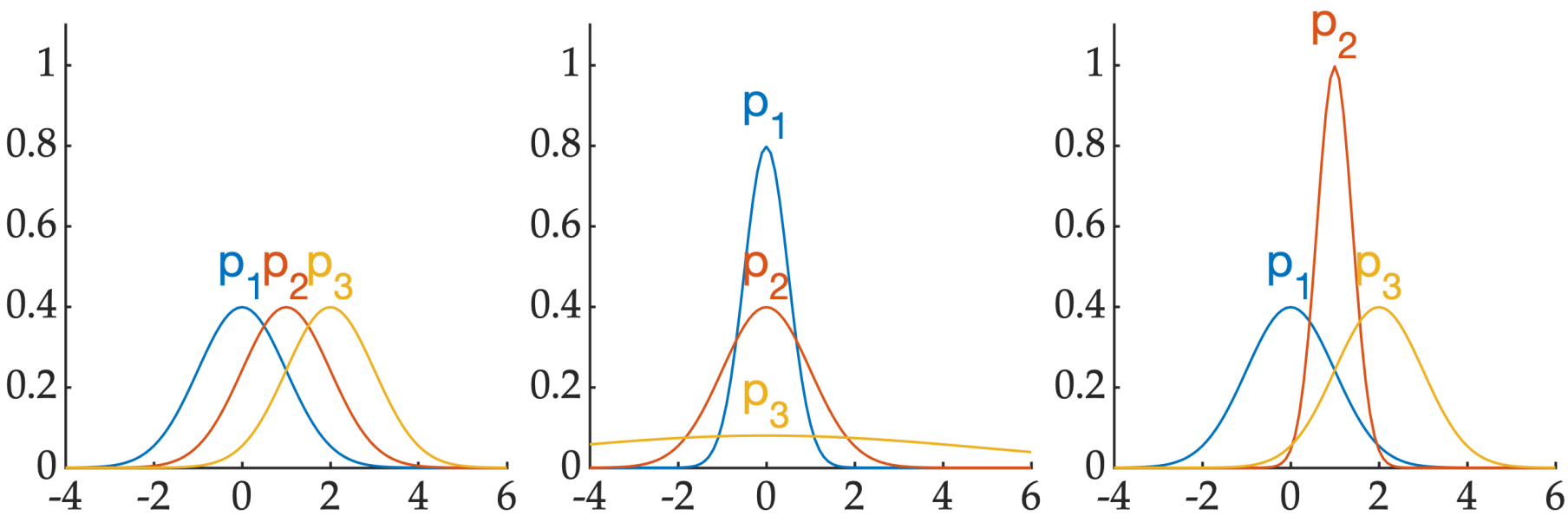
- Energy distance can be easily estimated from observations
- Equivalent to performing kernel archetypal analysis with

$$\kappa(p,q) = \mathbb{E}_{X\sim p,Y\sim q} k(X,Y) \text{ with } k(x,y) = ||x|| + ||y|| - ||x-y||.$$

Cons

- archetypes are interpretable as mixture

Archetypal Distribution Intuition



Partial Membership Models [Heller et al., 2008]

- For a mixture model with K mixture components,

$$p(\mathbf{x} \mid \Theta, \boldsymbol{\rho}) = \sum_{k=1}^{K} \rho_k p(\mathbf{x} \mid \boldsymbol{\theta}_k)$$

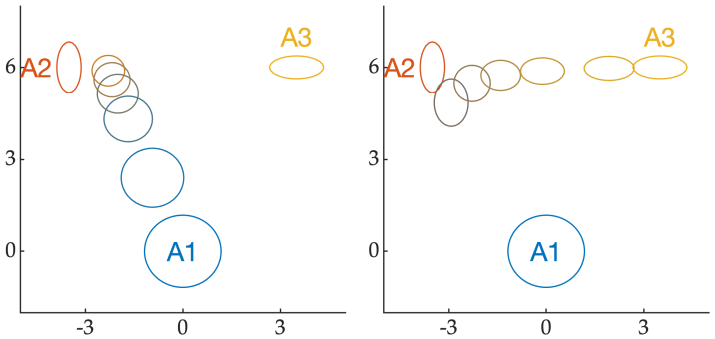
  where ρ_k are the mixing proportions.
- Given indicator variables $\boldsymbol{\zeta} = [\zeta_1, \ldots, \zeta_K]$ where $\zeta_k \in \{0, 1\}$ and $\sum_k \zeta_k = 1$,

$$p(\mathbf{x} \mid \Theta, \boldsymbol{\rho}) = \sum_{\boldsymbol{\zeta}} p(\boldsymbol{\zeta}) \prod_{k=1}^{K} p(\mathbf{x} \mid \boldsymbol{\theta}_k)^{\zeta_k}$$

  where $p(\ldots, \zeta_k = 1, \ldots) = \rho_k$.
- Partial membership model relaxes the constraint $\zeta_k \in \{0, 1\}$ to $\zeta_k \in [0, 1]$

$$p(\mathbf{x} \mid \Theta, \boldsymbol{\rho}) = \int p(\boldsymbol{\zeta} \mid \boldsymbol{\rho}) \left[\frac{1}{C(\boldsymbol{\zeta}, \Theta)} \prod_{k=1}^{K} p(\mathbf{x} \mid \boldsymbol{\theta}_k)^{\zeta_k} \right] \mathrm{d}\boldsymbol{\zeta}$$

  where C is a normalizing constant, and $p(\boldsymbol{\zeta} \mid \boldsymbol{\rho})$ is a distribution over simplex.



Exponential Family Distribution

- $\mathbf{x}$ is exponential family distributed if

$$\text{ExpFam}(\mathbf{x} \mid \boldsymbol{\theta}) = \exp(\mathbf{T}(\mathbf{x})^\intercal \boldsymbol{\eta}(\boldsymbol{\theta}) - A(\boldsymbol{\theta}))h(\mathbf{x})$$

  where $\boldsymbol{\eta}$ are the natural parameters, and $\mathbf{T}$ are the sufficient statistics
- For exponential family distributions, the partial membership product is also an exponential family distribution in the same family with natural parameters

$$\hat{\boldsymbol{\eta}} = \sum_k \zeta_k \boldsymbol{\eta}(\theta_k).$$

- For multivariate normal distribution,

$$\frac{1}{(2\pi)^{d/2}\det(\boldsymbol{\Sigma})^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right).$$

$$\boldsymbol{\eta} = [\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}; \text{vec}(\Sigma^{-1})] \text{ and } \mathbf{T} = [\mathbf{x}; \text{vec}(\mathbf{x}\mathbf{x}^\top)]$$

  where vec denotes the vectorization operator, and $[\cdot\,;\,\cdot]$ denotes column-wise concatenation.
- Therefore, partial membership product is a multivariate normal distribution with

$$\hat{\boldsymbol{\Sigma}}^{-1} = \sum \zeta_i {\boldsymbol{\Sigma}_i}^{-1} \text{ and } \hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\mu}} = \sum_i \zeta_i {\boldsymbol{\Sigma}_i}^{-1}\boldsymbol{\mu}_i.$$

Archetypal Analysis over Distributions

Given n distributions $p_1(\mathbf{x}), \ldots, p_n(\mathbf{x})$ on $\mathbb{R}^d$,

1. Define archetypal distributions $q_1(\mathbf{x}), \ldots, q_K(\mathbf{x})$ as

$$q_k(\mathbf{x}) = \frac{1}{c_k^w} \prod_{i=1}^{n} p_i(\mathbf{x})^{w_{ik}}$$

   where $\mathbf{W} \in \Delta$.
2. Reconstruct the distributions as

$$\hat{p}_j(\mathbf{x}) = \frac{1}{c_j^h} \prod_{k=1}^{K} q_k(\mathbf{x})^{h_{kj}}$$

   where $\mathbf{H} \in \Delta$.
3. Optimize the parameters $\mathbf{W}$ and $\mathbf{H}$ by minimizing

$$\sum_j D(\hat{p}_j(\mathbf{x}) \,\|\, p_j(\mathbf{x}))$$

   where D is a suitable divergence measure between two distributions.
4. Thus, archetypal analysis over distributions can be summarized as

$$\min_{\mathbf{W},\mathbf{H} \in \Delta} \sum_j D_{\mathrm{KL}} \left(\frac{1}{c_j} \prod_i p_i(\mathbf{x})^{\sum_k w_{ik} h_{kj}} \,\|\, p_j(\mathbf{x}) \right)$$

   where c_j is the respective normalizing constant.

Example

We solve this for multivariate normal distribution with diagonal covariance.

Archetypal distribution over 20 random bivariate normal distributions

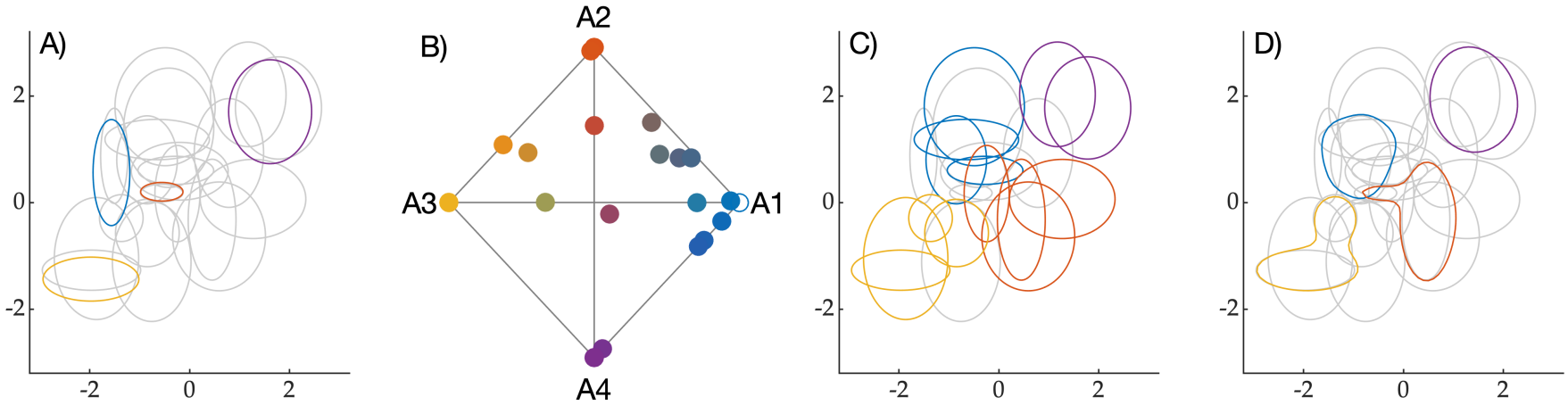


Figure: The figure compares several methods of archetypal analysis over a set of bivariate normal distributions (**grey**). The inferred archetypal distributions have been **color-coded**. (**A**) Archetypal analysis using the proposed approach, and (**B**) the resulting simplex plot. (**C**) Archetypal analysis using kernel archetypal analysis. (**D**) Archetypal analysis using statistical archetypal analysis.

Student Scores Data [Rovira et al., 2017]

Academic grades of 287 students over 68 total courses with missing values

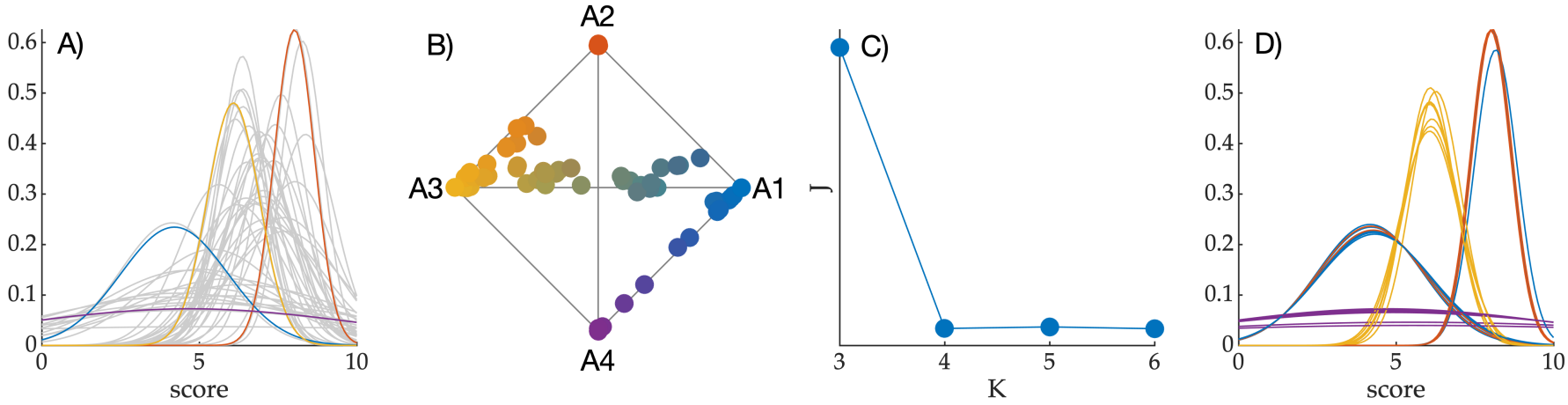


Figure: The figure shows the analysis of the student score dataset. (**A**) distributions of 100 student grades (**grey**) and the inferred archetypal distributions (**color-coded**), (**B**) the resulting simplex plot (**C**) the resulting cost values over different number of archetypes (**D**) inferred archetypes over 10 different random subsample of the data

1. A2) students who get consistently high grades and they are rare,
2. A3) students who get consistently moderate grades,
3. A1) students who get average grades and show more variability, and
4. A4) students who show high variability in their performance.

OpenML Data [Vanschoren et al., 2014]

Performance of 67 algorithms on 350 datasets

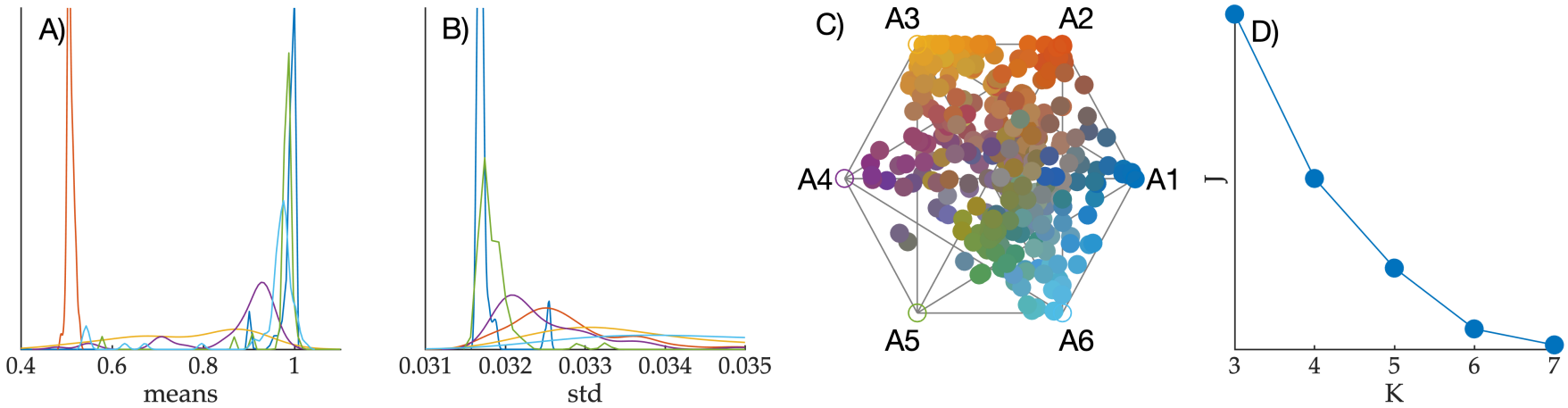


Figure: The figure shows the analysis of the algorithm dataset matrix. Distribution of (**A**) means and (**B**) variances respectively of 6 inferred archetypal distributions. Each distribution is estimated from 67 points since each archetype is a 67 dimensional distribution (**color-coded**). (**C**) The resulting simplex plot and (**D**) the resulting cost values over different number of archetypes.

1. A2) dataset where all algorithms perform at a chance level,
2. A1) dataset where all algorithms perform with the highest accuracy, and almost no variance
3. A3) dataset where the performance of algorithms show significant variability.
4. A5-A6) dataset where all algorithms perform very well but they show variation in different cross-validation folds with A6 showing more variation than A5.
5. A4) dataset where all algorithms perform moderately well, and also show some variation over validation folds.

Discussion

- Proposed an extension of archetypal analysis over distributions
- More interpretable than kernel and statistical archetypal analysis
- Complements interval archetypal analysis

- Only available for multivariate normal
- Requires better initialization and optimization
- Needs validation and interesting application

Thank you!

Reference

Adele Cutler and Leo Breiman. Archetypal analysis. *Technometrics*, 36(4):338–347, November 1994.

Maria R. D'Esposito, Francesco Palumbo, and Giancarlo Ragozini. Interval Archetypes: A New Tool for Interval Data Analysis. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 5(4):322–335, August 2012. ISSN 1932-1864. doi: 10.1002/sam.11140. URL `https://onlinelibrary.wiley.com/doi/full/10.1002/sam.11140`.

Katherine A. Heller, Sinead Williamson, and Zoubin Ghahramani. Statistical models for partial membership. In *Proceedings of the 25th international conference on Machine learning - ICML '08*, pages 392–399, Helsinki, Finland, 2008. ACM Press. ISBN 978-1-60558-205-4. doi: 10.1145/1390156.1390206. URL `http://portal.acm.org/citation.cfm?doid=1390156.1390206`.

Morten Mørup and Lars Kai Hansen. Archetypal analysis for machine learning and data mining. *Neurocomputing*, 80:54–63, March 2012. ISSN 0925-2312. doi: 10.1016/j.neucom.2011.06.033. URL `http://www.sciencedirect.com/science/article/pii/S0925231211006060`.

Sergi Rovira, Eloi Puertas, and Laura Igual. Data-driven system to predict academic grades and dropout. In *PLOS One*, 2017. URL `https://doi.org/10.1371/journal.pone.0171207`.

Sohan Seth and Manuel J. A. Eugster. Probabilistic archetypal analysis. *Machine Learning*, 102(1): 85–113, January 2016. ISSN 1573-0565. doi: 10.1007/s10994-015-5498-8. URL `https://doi.org/10.1007/s10994-015-5498-8`.

Joaquin Vanschoren, Jan N. van Rijn, Bernd Bischl, and Luis Torgo. OpenML: Networked Science in Machine Learning. *SIGKDD Explor. Newsl.*, 15(2):49–60, June 2014. ISSN 1931-0145. doi: 10.1145/2641190.2641198. URL `http://doi.acm.org/10.1145/2641190.2641198`.

Chenyue Wu and Esteban G Tabak. Statistical Archetypal Analysis. page 24, 2017.